

# DD2434/FDD3434 Machine Learning, Advanced Course

## Exercises (DGM, EM & VI)

Antonio Matosevic, Hazal Koptagel, Negar Safinianaini and Jens Lagergren \*

November 2019

### Abstract

This document consists of the theory, exercise questions and solutions for DD2434/FDD3434 Machine Learning, Advanced Course. Some of the solutions are taken from previous years' TAs; Borja Rodriguez Galvez and Elin Samuelsson.

## Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Directed Graphical Models (DGM) – Theory</b>                      | <b>2</b>  |
| 1.1      | Method 1 . . . . .   | 2         |
| 1.1.1    | Example . . . . .  | 3         |
| 1.2      | Method 2 . . . . .   | 4         |
| 1.2.1    | Example . . . . .  | 4         |
| <b>2</b> | <b>Directed Graphical Models (DGM) – Exercises</b>                   | <b>5</b>  |
| 2.1      | Bayes Ball . . . . .   | 5         |
| 2.2      | Bayes nets for a rainy day (Exercise 10.5 from Murphy [2]) . . . . . | 7         |
| <b>3</b> | <b>Expectation-Maximization – Theory</b>                             | <b>12</b> |
| <b>4</b> | <b>Expectation-Maximization – Exercises</b>                          | <b>14</b> |
| 4.1      | Mixture of scale mixtures . . . . .                                  | 14        |
| <b>5</b> | <b>Variational Inference – Theory</b>                                | <b>18</b> |
| 5.1      | Motivation . . . . .   | 18        |
| 5.2      | Main idea . . . . .  | 18        |
| <b>6</b> | <b>Variational Inference – Exercises</b>                             | <b>19</b> |

---

\*KTH Royal Institute of Technology, Stockholm, Sweden

# 1 Directed Graphical Models (DGM) – Theory

**D-separation rules:**

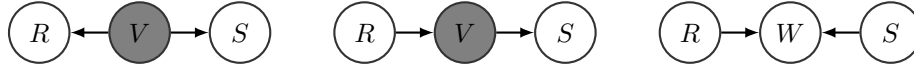


Figure 1: S and R are d-separated. Note that W is not observed and no descendants of W is observed either.

[Pearl 1988] example:

d-separated: dead battery  $\rightarrow$  car won't start  $\leftarrow$  no gas

not d-separated: dead battery  $\rightarrow$  **car won't start**  $\leftarrow$  no gas

**Two methods to determine if variables are independent or not:**

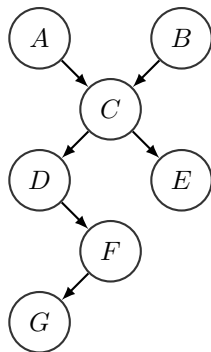
## 1.1 Method 1

This method is efficient when the dependencies are being examined among certain variables and not all the variables in the graph. We follow this procedure:

- 1. Construct the “ancestral graph” of all variables mentioned in the probability expression. This is a reduced version of the original net, consisting only of the variables **mentioned** and all of their **ancestors** (parents, parents’ parents, etc.)
- 2. “Moralize” the ancestral graph by “marrying” the parents. For each pair of variables with a common child, draw an undirected edge (line) between them. (If a variable has more than two parents, draw lines between every pair of parents.)
- 3. “Disorient” the graph by replacing the directed edges (arrows) with undirected edges (lines).
- 4. Delete the givens and their edges. If the independence question had any given variables, erase those variables from the graph and erase all of their connections, too.
- 5. Read the answer off the graph.
  - If the variables are disconnected in this graph, they are guaranteed to be independent.
  - If the variables are connected in this graph, they are not guaranteed to be independent.\* Note that “are connected” means “have a path between them,” so if we have a path X-Y-Z, X and Z are considered to be connected, even if there’s no edge between them.
  - If one or both of the variables are missing (because they were givens, and were therefore deleted), they are independent.

### 1.1.1 Example

Q: Are A and B conditionally independent, given D and F?



A: No. A and B are connected hence not conditionally independent, using method 1 as below:

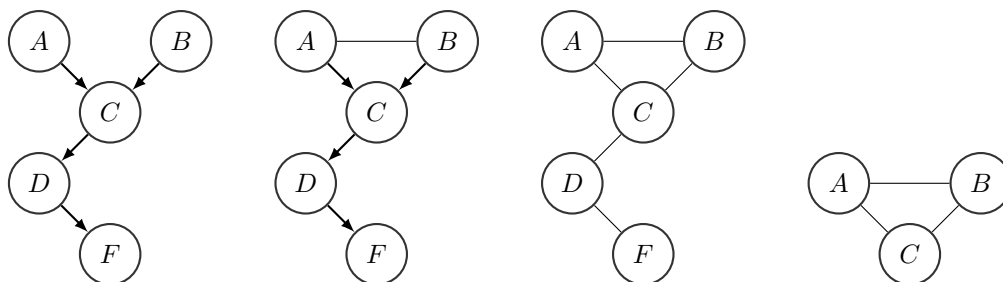


Figure 2: From left to right: 1. Draw ancestral graph 2. Moralize 3. Disorient 4. Delete givens

## 1.2 Method 2

This method is efficient when examining the dependence over the whole graph. We put a block symbol on the graph wherever we detect the d-separation using the rules in Figure 1. Then we can answer whether variables are d-separated or not by following the path between them i.e. if the path contains a block they are obviously d-separated.

### 1.2.1 Example

The blocks are illustrated as red point over the graph in Figure 3; the table shows the answers to d-separation which is gained by following the path between each pair of  $i$  and  $j$ .

| $i, j$ | d-separated |
|--------|-------------|
| A, D   | yes         |
| A, B   | no          |
| D, G   | yes         |
| D, H   | no          |
| D, E   | yes         |
| D, F   | no          |
| D, F   | no          |
| B, H   | yes         |

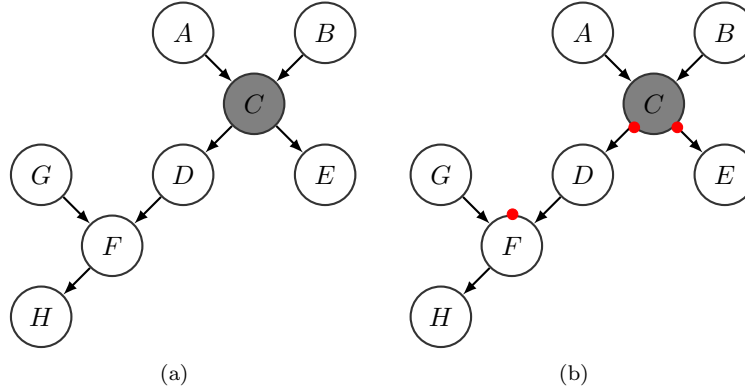


Figure 3: Blocks are shown in red on the graph in (b).

## 2 Directed Graphical Models (DGM) – Exercises

### 2.1 Bayes Ball

**Question:** List all variables that are independent of A given evidence on the shaded node for each of the DGMs a), b) and c) below.

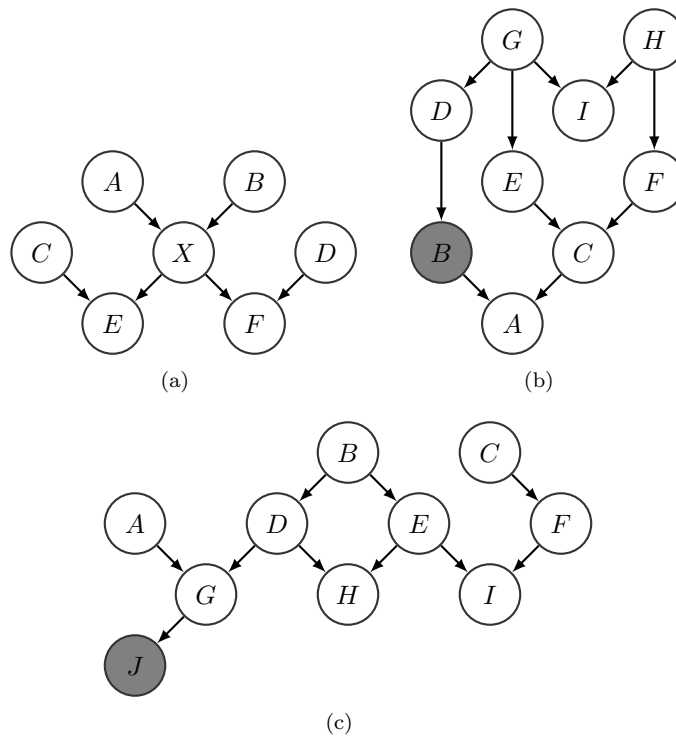


Figure 4: Some DGMs.

**Solution:** a) See Figure 5, b) See Figure 6, c) See Figure 7.

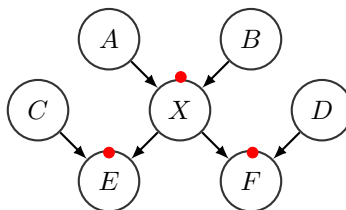


Figure 5: There is no conditioning on any node. We use method 2, So we just look for head-to-heads (the right most rule in Figure 1). Putting the blocks between the head-to-heads, we see that B,C, and D are separated from A or independent of A.

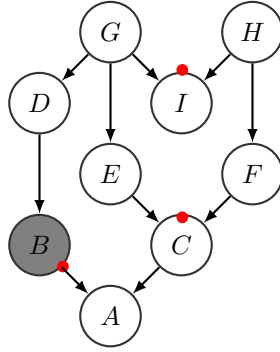


Figure 6: Following method 2, there is a path to A from all variables hence no variable is independent of A.

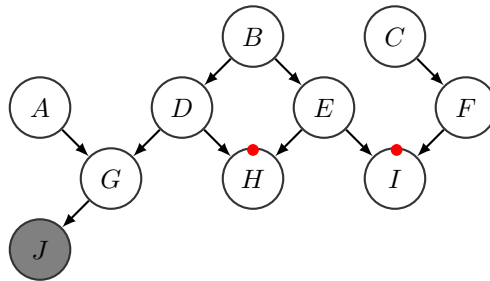


Figure 7: Following method 2, there is a path to A from all variables except F and C. Therefore, only C and F are independent of A.

## 2.2 Bayes nets for a rainy day (Exercise 10.5 from Murphy [2])

**Question:** (Source: Nando de Freitas) In this question you must model a problem with 4 binary variables:  $G = \text{"gray"}$ ,  $V = \text{"Vancouver"}$ ,  $R = \text{"rain"}$  and  $S = \text{"sad"}$ . Consider the directed graphical model describing the relationship between these variables shown in Figure 8 (and the probability tables shown in Table 1).

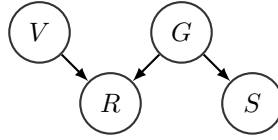


Figure 8: Bayesian net for a rainy day

Table 1: Probability tables of Bayes net for a rainy day

| $V = 0$  | $V = 1$      |
|----------|--------------|
| $\delta$ | $1 - \delta$ |

| $G = 0$  | $G = 1$      |
|----------|--------------|
| $\alpha$ | $1 - \alpha$ |

|         | $S = 0$  | $S = 1$      |
|---------|----------|--------------|
| $G = 0$ | $\gamma$ | $1 - \gamma$ |
| $G = 1$ | $\beta$  | $1 - \beta$  |

|           | $R = 0$ | $R = 1$ |
|-----------|---------|---------|
| $VG = 00$ | 0.6     | 0.4     |
| $VG = 01$ | 0.3     | 0.7     |
| $VG = 10$ | 0.2     | 0.8     |
| $VG = 11$ | 0.1     | 0.9     |

- Write down an expression for  $P(S = 1|V = 1)$  in terms of  $\alpha, \beta, \gamma, \delta$ .
- Write down an expression for  $P(S = 1|V = 0)$ . Is this the same or different to  $P(S = 1|V = 1)$ ? Explain why.
- Find maximum likelihood estimates of  $\alpha, \beta, \gamma$  using the following data set, where each row is a training case. (You may state your answers without proof.)

| $V$ | $G$ | $R$ | $S$ |
|-----|-----|-----|-----|
| 1   | 1   | 1   | 1   |
| 1   | 1   | 0   | 1   |
| 1   | 0   | 0   | 0   |

**Solution:** (The solution is taken from Elin Samuelsson's notes from December 2018 and modified slightly.)

- a. Write down an expression for  $P(S = 1|V = 1)$  in terms of  $\alpha, \beta, \gamma, \delta$ .

$$\begin{aligned}
P(S = 1|V = 1) &= \frac{P(S = 1, V = 1)}{P(V = 1)} \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^1 \sum_{r=0}^1 P(S = 1, V = 1, R = r, G = g) \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^1 \sum_{r=0}^1 P(S = 1, V = 1, R = r|G = g)P(G = g) \\
&\quad \{G \text{ is tail-to-tail and blocks the path } \rightarrow S \perp\!\!\!\perp \{R, V\}|G\} \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \sum_{r=0}^1 P(V = 1, R = r|G = g) \\
&= \frac{1}{P(V = 1)} \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \sum_{r=0}^1 P(R = r|V = 1, G = g)P(V = 1) \\
&= \frac{P(V = 1)}{P(V = 1)} \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \sum_{r=0}^1 P(R = r|V = 1, G = g) \\
&= \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \sum_{r=0}^1 P(R = r|V = 1, G = g) \\
&\quad \left\{ \sum_{r=0}^1 P(R = r|V = 1, G = g) = 1, \text{ regardless of the value of } G \text{ and } V \right\} \\
&= \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \\
&= P(S = 1|G = 0)P(G = 0) + P(S = 1|G = 1)P(G = 1) \\
&= \alpha(1 - \gamma) + (1 - \alpha)(1 - \beta) \\
&= 1 - \beta + \alpha\beta - \alpha\gamma
\end{aligned} \tag{1}$$

- b. Write down an expression for  $P(S = 1|V = 0)$ . Is this the same or different to  $P(S = 1|V = 1)$ ? Explain why.

$$\begin{aligned}
P(S = 1|V = 0) &= \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \sum_{r=0}^1 P(R = r|V = 0, G = g) \\
&= \sum_{g=0}^1 P(S = 1|G = g)P(G = g) \\
&= P(S = 1|V = 1)
\end{aligned} \tag{2}$$

since  $\sum_{r=0}^1 P(R = r|V = v, G = g) = 1$ , regardless of the value of  $G$  and  $V$ .

- c. Find maximum likelihood estimates of  $\alpha, \beta, \gamma$  using the following data set, where each row is a training case. (You may state your answers without proof.)



| $V$ | $G$ | $R$ | $S$ |
|-----|-----|-----|-----|
| 1   | 1   | 1   | 1   |
| 1   | 1   | 0   | 1   |
| 1   | 0   | 0   | 0   |

Notation:

$$\mathbf{1}(a = b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b \end{cases} \quad (3)$$

Maximum likelihood estimates:

$$\hat{\alpha} = P(G = 0) = \frac{\sum_{n=1}^N \mathbf{1}(G_n = 0)}{N} = \frac{1}{3} \quad (4)$$

$$\hat{\beta} = P(S = 0|G = 1) = \frac{P(S = 0, G = 1)}{P(G = 1)} = \frac{\sum_{n=1}^N \mathbf{1}(S_n = 0, G_n = 1)}{\sum_{n=1}^N \mathbf{1}(G_n = 1)} = \frac{0}{2} \quad (5)$$

$$\hat{\gamma} = P(S = 0|G = 0) = \frac{P(S = 0, G = 0)}{P(G = 0)} = \frac{\sum_{n=1}^N \mathbf{1}(S_n = 0, G_n = 0)}{\sum_{n=1}^N \mathbf{1}(G_n = 0)} = \frac{1}{1} \quad (6)$$

### An Alternative Solution

The more general way to find the MLE is i) write the likelihood (or log-likelihood) ii) take derivative w.r.t the parameter of interest and set it to zero iii) check whether the value actually maximizes the likelihood (by looking at the second derivative [3] is negative or not).

The likelihood of the data is:

$$\begin{aligned} \mathcal{L} &= P(D|\Theta) \\ &= P(D_1, \dots, D_N|\Theta) \\ &= \prod_{n=1}^N P(D_n|\Theta) \\ &= \prod_{n=1}^N P(V_n|\delta)P(G_n|\alpha)P(S_n|G_n, \beta, \gamma)P(R_n|V_n, G_n) \\ &= P(V_n = 1|\delta)^3 P(G_n = 0|\alpha)P(G_n = 1|\alpha)^2 P(S_n = 0|G_n = 0, \beta, \gamma)P(S_n = 1|G_n = 1, \beta, \gamma)^2 \\ &\quad P(R_n = 0|V_n = 1, G_n = 0)P(R_n = 0|V_n = 1, G_n = 1)P(R_n = 1|V_n = 1, G_n = 1) \\ &= (1 - \delta)^3 \alpha (1 - \alpha)^2 \gamma (1 - \beta)^2 \times 0.2 \times 0.1 \times 0.9 \\ &\propto (1 - \delta)^3 \alpha (1 - \alpha)^2 \gamma (1 - \beta)^2 \end{aligned} \quad (7)$$

Let's look at the likelihood (see Figure 9). For the first subplot, I fixed  $\alpha, \gamma, \beta \in (0, 1)$  and ranged  $\delta \in [0, 1]$ . The subplot shows how the likelihood changes w.r.t  $\delta$ . I repeated the same method for the rest of the parameters. From the figure, we can clearly see which values of the parameters maximize the likelihood ( $\hat{\delta} = 0, \hat{\alpha} = 0.33, \hat{\gamma} = 1, \hat{\beta} = 0$ ).

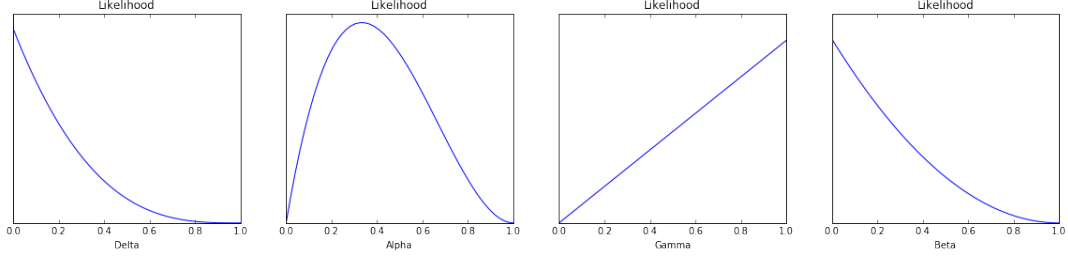


Figure 9: Likelihood of rainy day example with varying parameters

Now, let's show these with the derivatives. First, consider  $\alpha$ . Take the first derivative of the likelihood w.r.t  $\alpha$ .

$$\begin{aligned}
\frac{\partial}{\partial \alpha} \mathcal{L} &= \frac{\partial}{\partial \alpha} \alpha(1-\alpha)^2 C \\
&= (1-\alpha)^2 C - 2\alpha(1-\alpha)C \\
&= (1-\alpha)(1-\alpha-2\alpha)C \\
&= (1-\alpha)(1-3\alpha)C \\
&= 0 \\
\alpha &= 1 \text{ or } \alpha = \frac{1}{3}
\end{aligned} \tag{8}$$

where  $C$  is a temporary variable that I used to represent all the other terms in the likelihood beside the parameter of interest (notice that  $C$  is non-negative). There are two  $\alpha$  values which either maximize or minimize the likelihood. We need to check the second derivative of the likelihood w.r.t  $\alpha$ :

$$\begin{aligned}
\frac{\partial^2}{\partial \alpha^2} \mathcal{L} &= \frac{\partial}{\partial \alpha} \alpha(1-\alpha)^2 C \\
&= -(1-3\alpha)C - 3(1-\alpha)C \\
&= (-4+6\alpha)C
\end{aligned} \tag{9}$$

When  $\alpha = 1$ , the second derivative becomes  $2C$ , which is non-negative, which means  $\alpha = 1$  is not a maximizer of the likelihood. When  $\alpha = \frac{1}{3}$ , the second derivative becomes  $-2C$ , which is negative, which means  $\hat{\alpha} = \frac{1}{3}$  is the maximum likelihood estimator. We can confirm this result with Figure 9.

Now, we move on to  $\beta$ . We re-write the likelihood as  $\mathcal{L} = (1-\beta)^2 C$ . It is clear that the  $\hat{\beta}$  which maximizes the likelihood must be  $\hat{\beta} = 0$  (since  $C$  is non-negative and  $\mathcal{L}$  gets the highest value, which is  $1C$ , when  $\beta = 0$ ). Let's look at take the derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \beta} \mathcal{L} &= \frac{\partial}{\partial \beta} (1-\beta)^2 C \\
&= -2(1-\beta)C \\
&= 0 \\
\beta &= 1
\end{aligned} \tag{10}$$

Now, check the second derivative:

$$\begin{aligned}\frac{\partial^2}{\partial \beta^2} \mathcal{L} &= \frac{\partial}{\partial \alpha} (1 - \beta)^2 C \\ &= 2C \\ &> 0\end{aligned}\tag{11}$$

Since the second derivative is always non-negative,  $\beta = 1$  is the minimizer of the likelihood. Notice that we were unable to find the  $\beta$  which maximizes the likelihood with this approach. Why? In our data  $D$ , we don't have any samples where  $S = 0$  and  $G = 1$ .

Finally, let's re-write the likelihood in terms of  $\gamma$ ;  $\mathcal{L} = \gamma C$ . It is clear that the  $\hat{\gamma}$  which maximizes the likelihood must be  $\hat{\gamma} = 1$  (because the likelihood is a linearly increasing function of  $\gamma$ ).

### 3 Expectation-Maximization – Theory

**General setup** Let  $\mathbf{x} = \{x_1, \dots, x_N\}$  be a set of observations (of corresponding random variables (RVs)  $X_1, \dots, X_N$ ) and  $\mathbf{Z} = \{Z_1, \dots, Z_K\}$  a set of latent RVs. Oftentimes one constructs a model where  $\mathbf{x}$  and  $\mathbf{Z}$  are directly connected and parameterises it (i.e. distributions involved) by a set of parameters  $\theta$ . The goal is then to infer  $\theta$  given data  $\mathbf{x}$ .

**Maximum likelihood estimation** A common tool for estimating distribution parameters is the maximum likelihood estimation (MLE). In its essence it seeks to find the point estimate  $\theta^*$  of  $\theta$  so that the probability density of observing exactly this sample  $\mathbf{x}$  of  $X_1, \dots, X_n$  is maximized. This probability density is referred to as data likelihood and denoted as  $p(\mathbf{x}; \theta)$ , which is just uncluttered way of writing  $p(X_1 = x_1, \dots, X_N = x_n; \theta)$ , henceforth used for all distributions.

**MLE for latent variables models** In the context of latent variable models, data likelihood can be written using the law of total probability as

$$p(\mathbf{x}; \theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta), \quad (12)$$

for **all** possible combinations  $\mathbf{z} = \{z_1, \dots, z_K\}$  that  $Z_1, \dots, Z_K$  can attain. It is obvious that this sum<sup>1</sup> is computationally exponential, often referred to as intractable<sup>2</sup>.

**Expectation-maximization algorithm** To alleviate this problem, one can optimize something close to likelihood, but tractable. Let us present one way to do this by maximizing its lower bound instead. Returning to Eq. 12 we can, thanks to monotonicity, equivalently maximize data likelihood in the log-space i.e. we have that

$$\log p(\mathbf{x}; \theta) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta). \quad (13)$$

Now, there is a result called Jensen’s inequality stating that for a concave function  $f(\cdot)$  of a RV  $\Psi$  it holds that  $f(\mathbb{E}[\Psi]) \geq \mathbb{E}[f(\Psi)]$ . Since the logarithm conveniently appearing in Eq. 13 is concave, we can perhaps use this inequality to somehow simplify the initial problem. It only remains to rewrite Eq. 13 into an expectation so that the inequality can be applied. The trick is a common mathematical manipulation of multiplying and dividing by the same thing, essentially not changing the expression, but making it possible to manipulate the expression

<sup>1</sup>Or integral in case of continuous latent RVs.

<sup>2</sup>This terms also applies to integrals without closed form in the continuous case.

in a favourable way. Following this, we can write

$$\log p(\mathbf{x}; \theta) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta) \quad (14)$$

$$= \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta) \frac{q(\mathbf{z})}{q(\mathbf{z})} \quad (15)$$

$$= \log \mathbb{E}_q \left[ \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \right] \quad (16)$$

$$\geq \mathbb{E}_q \left[ \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \right] \quad (17)$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})}, \quad (18)$$

for any  $q(\mathbf{z})$ . Now that we have a lower bound to the true log-likelihood, it remains to choose  $q(\mathbf{z})$  wisely so that the lower bound is as tight as possible. By rewriting Eq. 18 as

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} = -\mathbb{KL}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}; \theta)) + \log p(\mathbf{x}; \theta) \quad (19)$$

one can notice (ignoring the second term since it does not depend on  $q(\mathbf{z})$ ) that the inequality becomes equality for  $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \theta)$ . However,  $\theta$  parameterising this  $q(\mathbf{z})$  is exactly what we wanted to estimate from the very beginning and now we need it to compute the lower bound, which we need to estimate  $\theta$ , which we need to compute lower bound etc. This cyclic reasoning implies an iterative algorithm that will at one stage (E-step) re-estimate

$$q^t = \arg \max_q \mathcal{L}(q, \theta^t),$$

where  $\theta^t$  is fixed and comes from the previous iteration<sup>3</sup>, and at the next stage (M-step) re-estimate

$$\theta^{t+1} = \arg \max_{\theta} \mathcal{L}(q^t, \theta),$$

where now  $q^t$  is fixed and comes from the current iteration. This algorithm is better known as expectation-maximization (EM), although it is here actually presented as maximization-maximization.

*Note:* Now that we have introduced an iterative scheme, we simply have that  $q^t = p(\mathbf{z}|\mathbf{x}; \theta^t)$ , which can somewhat simplify the expression for  $\mathcal{L}(q, \theta)$  at the M-step. By plugging  $q^t$  in  $\mathcal{L}(q, \theta)$ , one can easily show that

$$\mathcal{L}(q^t, \theta) = Q(\theta, \theta^t) + c, \quad (20)$$

where  $Q(\theta, \theta^t) = \mathbb{E}_{q^t}[\log p(\mathbf{x}, \mathbf{z}; \theta)]$ , also called expected complete log-likelihood<sup>4</sup>, and  $c$  is the entropy of  $q^t$  which is constant w.r.t.  $\theta$ <sup>5</sup>. Hence, it is actually enough to maximize only  $Q(\theta, \theta^t)$  with respect to  $\theta$  at the M-step instead of the whole  $\mathcal{L}(q^t, \theta)$ .

---

<sup>3</sup>At the first iteration we initialize  $\theta^t$ .

<sup>4</sup>It is this expectation that motivates the name E-step.

<sup>5</sup>Here it is crucial to distinguish between fixed  $\theta^t$  from the previous iteration that  $q^t$  depends on and  $\theta$  as a variable parameter to be maximized w.r.t.

## 4 Expectation-Maximization – Exercises

### 4.1 Mixture of scale mixtures

Usual Gaussian mixture model (GMM) is represented as a weighted sum of  $k$  components, each being a Gaussian distribution with some mean and variance. The goal in such model is to estimate the mixture weights, as well as all means and variances. One can extend this model to be more flexible by making each component a mixture on its own. Namely, instead of components being Gaussian distributions, we could instead model them as mixtures of Gaussians with the same mean, but different variances (so-called scale mixtures). Graphical model in plate notation for this setup is given in Figure 10.

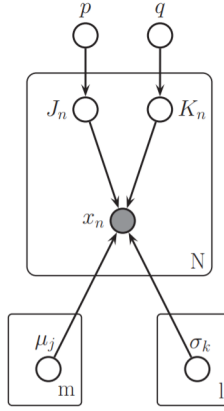


Figure 10: Mixture of scale mixtures

We can interpret  $J_n$  as the latent variable specifying which mean to use, and analogously  $K_n$  which variance to use. With this notation we can formally define the scale mixture as

$$p(x_n | J_n = j) = \sum_{k=1}^l q_k \mathcal{N}(\mu_j, \sigma_k^2), \quad (21)$$

where  $q_j := p(K_n = k)$ , and the mixture of scale mixtures as

$$p(x_n) = \sum_{j=1}^m p_j \left[ \sum_{k=1}^l q_k \mathcal{N}(\mu_j, \sigma_k^2) \right], \quad (22)$$

where  $p_j := p(J_n = j)$ , for all  $n$ .<sup>6</sup>

**Task 1.** Given the observations  $\mathbf{x} = \{x_n\}_{n=1}^N$  estimate using EM algorithm parameters  $\theta = \{p_1, \dots, p_m, \mu_1, \dots, \mu_m, q_1, \dots, q_l, \sigma_1^2, \dots, \sigma_l^2\}$ .

<sup>6</sup>Note that Eq. 21 and 22 are simple applications of marginalization and conditioning, e.g. for 21  $p(x_n | J_n = j) = \sum_{k=1}^l p(x_n, K_n = k | J_n = j) = \sum_{k=1}^l p(x_n | J_n = j, K_n = k) p(K_n = k)$ .

**Solution:**

Referring now to the theory part in Section 3, we recall having the E-step and the M-step. In the E-step, we were looking for the optimal  $q^t$ , that has been shown to be  $p(\mathbf{x}, \mathbf{z}; \theta^t)$ , here in particular for  $\mathbf{z} = \{\mathbf{J}, \mathbf{K}\}$ , where  $\mathbf{x} = \{x_n\}_{n=1}^N$ ,  $\mathbf{J} = \{J_n\}_{n=1}^N$  and  $\mathbf{K} = \{K_n\}_{n=1}^N$ . We can even somewhat simplify this term by observing that  $J_n \perp K_n$ . Omitting  $\theta^t$  to keep the notation uncluttered and assuming the i.i.d. data<sup>7</sup>, we have that

$$p(\mathbf{x}, \mathbf{J}, \mathbf{K}) = p(\mathbf{x}|\mathbf{J}, \mathbf{K})p(\mathbf{J})p(\mathbf{K}) \quad (23)$$

$$= \prod_{n=1}^N \{p(x_n|J_n, K_n)p(J_n)p(K_n)\} \quad (24)$$

$$= \prod_{n=1}^N \prod_{k=1}^l \prod_{j=1}^m \{p(x_n|J_n = j, K_n = k)p(J_n = j)p(K_n = k)\}^{\mathbf{1}_{(J_n=j, K_n=k)}}, \quad (25)$$

where  $\mathbf{1}_{(J_n = j, K_n = k)}$  is the indicator function specifying what mean and variance are being considered for a given  $x_n$  (since  $J_n$  and  $K_n$  are otherwise not observed)<sup>8</sup>. For simplicity of the expressions and to avoid numerical underflow in implementation, we can work in the log-space, where we have

$$\log p(\mathbf{x}, \mathbf{J}, \mathbf{K}) = \sum_{n=1}^N \sum_{k=1}^l \sum_{j=1}^m \mathbf{1}_{(J_n = j, K_n = k)} \{ \log p(\mathbf{x}_n|J_n = j, K_n = k) \quad (26)$$

$$+ \log p(J_n = j) + \log p(K_n = k) \}. \quad (27)$$

Now that we have optimal  $q^t$ , we can compute  $Q(\theta, \theta^t)$  that we showed is enough to maximize to obtain optimal  $\theta$  in the M-step. Namely, we have that

$$Q(\theta, \theta^t) = \mathbb{E}_{q^t}[\log p(\mathbf{x}, \mathbf{J}, \mathbf{K})] \quad (28)$$

$$= \sum_{n=1}^N \sum_{k=1}^l \sum_{j=1}^m \mathbb{E}_{q^t}[\mathbf{1}_{(J_n = j, K_n = k)}] \{ \log \mathcal{N}(\mu_j, \sigma_k^2) + \log p_j + \log q_k \} \quad (29)$$

where

$$\mathbb{E}_{q^t}[\mathbf{1}_{(J_n = j, K_n = k)}] = \sum_{j'=1}^m \sum_{k'=1}^l \mathbf{1}_{(J_n = j, K_n = k)} p(J_n = j', K_n = k'|x_n, \theta^t) \quad (30)$$

$$= p(J_n = j, K_n = k|x_n, \theta^t) \quad (31)$$

$$= \frac{p(J_n = j, K_n = k, x_n|\theta^t)}{p(x_n|\theta^t)} \quad (32)$$

$$= \frac{p_j^t q_k^t \mathcal{N}(\mu_j^t, (\sigma_k^t)^2)}{\sum_{j'=1}^m \sum_{k'=1}^l p_{j'}^t q_{k'}^t \mathcal{N}(\mu_{j'}^t, (\sigma_{k'}^t)^2)}, \quad (33)$$

where the superscript  $^t$  denotes that these values are parameter estimates from the previous ( $t$ -th) iteration, reflecting that the expectation is taken w.r.t  $q^t$ . Let us also for convenience

<sup>7</sup>Formally, it is assumed that random variables  $X_1, \dots, X_N$ , whose realizations are  $x_1, \dots, x_N$ , are i.i.d.

<sup>8</sup>Note that now the probability densities can simply be evaluated, and the expectation needed to compute  $Q$  is acting only on the indicator.

of notation denote  $\gamma_{j,k,n} := \mathbb{E}_{q^t}[\mathbb{1}(J_n = j, K_n = k)]$ . Finally, by unraveling  $\log \mathcal{N}(\mu_j, \sigma_k^2)$  the  $Q$ -function can be explicitly written as

$$Q(\theta, \theta^t) = \sum_{n=1}^N \sum_{k=1}^l \sum_{j=1}^m \gamma_{j,k,n} \left( -\frac{1}{2} \log \sigma_k^2 - \frac{1}{2} \log(2\pi) - \frac{1}{2} \frac{(x_n - \mu_j)^2}{\sigma_k^2} + \log p_j + \log q_k \right) \quad (34)$$

Now we can find the update expressions for the parameters (i.e. perform the M-step) by maximizing the obtained  $Q$ -function. To find the parameter  $\mu_j$  and  $\sigma_k^2$  maximizing  $Q$ , one can take the corresponding partial derivatives and solve  $\partial_{\mu_j} Q = 0$  and  $\partial_{\sigma_k^2} Q = 0$ . For  $p_j$  and  $q_k$ , one needs to be slightly more careful – since these parameters are in fact probabilities, one must satisfy the constraints that  $\sum_{j=1}^m p_j = 1$  and  $\sum_{k=1}^l q_k = 1$ . How to tackle this, we will show later on. Attentive reader<sup>9</sup> could have also notice that we cheated a bit here; namely one should actually jointly maximize  $Q$  w.r.t.  $\theta$ , and not w.r.t. each parameter separately keeping the others fixed. As it turns out, this is rather difficult, which is why it is a common practice to take the suggested approach, which is sometimes referred to as generalized EM. Let us first find the update expression for  $\mu_j$ . Note that  $j$  is fixed here, hence no summation over it, so we have

$$\partial_{\mu_j} Q \stackrel{\pm}{=} \partial_{\mu_j} \sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \left( -\frac{(x_n - \mu_j)^2}{2\sigma_k^2} \right) = \sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \frac{x_n - \mu_j}{\sigma_k^2} = 0 \quad (35)$$

$$\iff \sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \frac{x_n}{\sigma_k^2} = \mu_j \sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \frac{1}{\sigma_k^2} \quad (36)$$

$$\iff \mu_j = \frac{\sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \frac{x_n}{\sigma_k^2}}{\sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} \frac{1}{\sigma_k^2}}, \quad (37)$$

where we used  $\stackrel{\pm}{=}$  to denote equality up to an additive constant. Similarly, for  $\sigma_k^2$  we have

$$\partial_{\sigma_k^2} Q \stackrel{\pm}{=} \partial_{\sigma_k^2} \sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n} \left( -\frac{\log \sigma_k^2}{2} - \frac{(x_n - \mu_j)^2}{2\sigma_k^2} \right) = \sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n} \left( -\frac{1}{2\sigma_k^2} + \frac{(x_n - \mu_j)^2}{2(\sigma_k^2)^2} \right) = 0 \quad (38)$$

$$\iff \frac{1}{2\sigma_k^2} \sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n} = \frac{1}{2(\sigma_k^2)^2} \sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n} (x_n - \mu_j)^2 \quad (39)$$

$$\iff \sigma_k^2 = \frac{\sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n} (x_n - \mu_j)^2}{\sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n}}. \quad (40)$$

Now, as mentioned, to obtain the update equations for  $p_j$  and  $q_k$  we have to solve a constrained optimization problem, where the constraints are as given before. Such problems are usually approached using Lagrangian multipliers that convert the problem to unconstrained optimization where the new objective is  $\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$ , where  $f(x)$  is the old objective,  $g(x)$  constraint and  $\lambda$  the Lagrangian multiplier. Since  $\mathcal{L}$  depends on both  $x$  and  $\lambda$ , it is necessary to optimize w.r.t. to both, and this can again be done with partial derivatives.

<sup>9</sup><http://www.ma.rhul.ac.uk/~uvah099/Sat/reader.html>



Applying this to our setup, we have the objective for  $p_j$

$$\mathcal{L}(p_j, \lambda) = Q(\theta, \theta^t) - \lambda \left( 1 - \sum_{j'=1}^m p_{j'} \right) \quad (41)$$

$$\stackrel{+}{=} \sum_{n=1}^N \sum_{k=1}^l \{ \gamma_{j,k,n} \log p_j \} - \lambda \left( 1 - \sum_{j'=1}^m p_{j'} \right). \quad (42)$$

Firstly, the partial derivative w.r.t.  $p_j$  gives

$$\partial_{p_j} \mathcal{L}(p_j, \lambda) = \frac{1}{p_j} \sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n} - \lambda = 0 \quad (43)$$

$$p_j = \frac{\sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n}}{\lambda}. \quad (44)$$

We can see that the update of  $p_j$  depends on  $\lambda$ . To obtain the explicit expression for  $\lambda$ , we first take the partial derivative of  $\mathcal{L}$  w.r.t. it

$$\partial_{\lambda} \mathcal{L}(p_j, \lambda) = 1 - \sum_{j'=1}^m p_{j'} = 0 \quad (45)$$

$$\sum_{j'=1}^m p_{j'} = 1, \quad (46)$$

which can be used to obtain

$$\lambda = \lambda \cdot 1 = \lambda \sum_{j'=1}^m p_{j'} = \sum_{j'=1}^m p_{j'} \lambda = \sum_{j'=1}^m \sum_{n=1}^N \sum_{k=1}^l \gamma_{j',k,n}, \quad (47)$$

where the last equality comes from rearranging Equation 44. Hence, the update equation for  $p_j$  is given by

$$p_j = \frac{\sum_{n=1}^N \sum_{k=1}^l \gamma_{j,k,n}}{\sum_{j'=1}^m \sum_{n=1}^N \sum_{k=1}^l \gamma_{j',k,n}}. \quad (48)$$

One can in the exactly same manner obtain the update equation for  $q_k$ . For brevity we omit these calculations and conclude with the update equation

$$q_k = \frac{\sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k,n}}{\sum_{k'=1}^l \sum_{n=1}^N \sum_{j=1}^m \gamma_{j,k',n}}. \quad (49)$$

## 5 Variational Inference – Theory

A deterministic approximate inference algorithm.

### 5.1 Motivation

We can't compute the posterior for many interesting models. For example for the Bayesian mixture of Gaussian, we draw  $z_i \sim \text{Mult}(\pi)$  and  $x_i \sim N(\mu_{z_i}, \sigma^2)$  resulting in the following posterior [1]:

$$p(\mu_{1:K}, z_{1:n} | x_{1:n}) = \frac{\prod_k p(\mu_k) \prod_i p(z_i) p(x_i | z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_k p(\mu_k) \prod_i p(z_i) p(x_i | z_i, \mu_{1:K})} \quad (50)$$

The denominator is the problem; even if we take the summation outside, this is intractable when  $n$  is reasonably large.

### 5.2 Main idea

So far we used EM and did point estimate for the model parameters; now we want to find posterior distribution for the **unknown** model parameters and hidden variables. For a DGM with observations  $X$ , hidden variables  $Z$  and model parameters  $\Theta$ , want to pick an approximation  $q(Z, \Theta)$  to the distribution from some tractable family, and then to try to make this approximation as close as possible to the true posterior. This reduces inference to an optimization problem [2]. We measure the closeness of the two distributions  $q$  and  $p$  with Kullback-Leibler (KL) divergence.

$$\mathbb{KL}(q||p) = \sum_z q(Z, \Theta) \log \frac{q(Z, \Theta)}{p(Z, \Theta|X)} \quad (51)$$

$$\sum_z q(Z, \Theta) \log \frac{q(Z, \Theta)}{p(Z, \Theta|X)} = -E_q[\log p(X, Z, \Theta)] + E_q[\log q(Z, \Theta)] + \log p(X) \quad (52)$$

We cannot actually minimize KL divergence in Eq. 51 but since we have Eq. 52, we maximise lower bound of log likelihood, Evidence Lower Bound (ELBO) [1], which is (same as equation 18):

$$\text{ELBO}(q) = E_q[\log P(X, Z, \theta)] - E_q[\log q(Z, \theta)]. \quad (53)$$

Note that we can use ELBO for convergence test at each iteration i.e. the difference of the ELBO of current value and the previous one should be smaller than some small epsilon.

In mean field Variational Inference, we assume that the variational family factorizes,

$$q(Z_1, \dots, Z_n, \Theta_1, \dots, \Theta_k) = \prod_i q(Z_i) \prod_k q(\Theta_k) \quad (54)$$

Now for each update equation of  $q(z_i)$  we perform the expectation, over the log of the joint distribution, w.r.t. all the hidden variables except the one we are deriving the approximate posterior for. For example, we obtain the update equation for  $z_j$ ,  $q(z_j)$ , by calculating  $E_{-z_j}(\log P(X, Z, \Theta))$ . Here we explain [1] that this is the result of the maximization of ELBO. We rewrite the ELBO and then take the partial derivative and set it to zero.

If we call the set of latent variables and parameters,  $\Psi$ , and use the chain rule, we can write  $p(X, \Psi) = p(X) \prod_j p(\Psi_j | \Psi_{1:j-1}, X)$ . Now the ELBO becomes:

$$\text{ELBO}(q) = \log P(X) + \sum_j E[\log P(\Psi_j | \Psi_{1:j-1}, X)] - E_j[\log q(\Psi_j)] \quad (55)$$

Now if you write the ELBO as a function of the last variable in the chain, say  $\Psi_k$ , you get:

$$\mathcal{L}(q(\Psi_k)) = \text{const} + E[\log P(\Psi_k | \Psi_{-k}, X)] - E[\log q(\Psi_k)] = \quad (56)$$

$$\int q(\Psi_k) E_{-k}[\log P(\Psi_k | \Psi_{-k}, X)] d\Psi_k - \int q(\Psi_k) \log q(\Psi_k) d\Psi_k \quad (57)$$

Taking the partial derivative of the integrals in 57 and setting it to zero you get the update equation for each posterior during the coordinate ascent algorithm:

$$q^*(\Psi_k) \propto E_{-k}[\log P(\Psi_k | \Psi_{-k}, X)] \propto E_{-k}[\log P(\Psi, X)] \quad (58)$$

## 6 Variational Inference – Exercises

The Cartesian Matrix Model (CMM) is defined as follows. There are  $R$  row distributions  $\{N(\mu_r, \lambda_r^{-1}) : 1 \leq r \leq R\}$ , each variance  $\lambda_r^{-1}$  is known and each  $\mu_r$  has prior distribution  $N(\mu, \lambda^{-1})$ . There are also  $C$  column distributions  $\{N(\xi_c, \tau_c^{-1}) : 1 \leq c \leq C\}$ , each variance  $\tau_c^{-1}$  is known and each  $\xi_c$  has prior distribution  $N(\xi, \tau^{-1})$ . All hyper-parameters are known. A matrix  $S$  is generated by, for each row  $1 \leq r \leq R$  and each column  $1 \leq c \leq C$ , setting  $S_{rc} = X_r + Y_c$  where  $X_r$  is sampled from  $N(\mu_r, \lambda_r^{-1})$  and  $Y_c$  from  $N(\xi_c, \tau_c^{-1})$ . Use Variational Inference in order to obtain a variational distribution

$$q(\mu_1, \dots, \mu_R, \xi_1, \dots, \xi_C) = \prod_r q(\mu_r) \prod_c q(\xi_c)$$

that approximates  $p(\mu_1, \dots, \mu_R, \xi_1, \dots, \xi_C | S)$ . Tip: what distribution do you get from the sum of two Gaussian random variables? What is the relation between the means?

**Question 15:** *Present the algorithm written down in a formal manner (using both text and mathematical notation, but not pseudo code).*

Figure 11: From Assignment 2, 2017

**Question:** See the question 15 from Figure 11.

**Solution:**

Since each cell of the matrix denoted as  $S_{rc}$  is a sum of two Gaussians then as we know we can get the Gaussian result where the mean is the sum of the mean of  $X$  and the mean of  $Y$ . The variance is also summed. So by having  $\mu_r$  and  $\epsilon_c$ , the distribution of each cell can be expressed as  $\mathcal{N}(\mu_r + \epsilon_c, \lambda_r^{-1} + \tau_c^{-1})$ . Below are the calculations of  $q_l(\mu_l)$  and  $q_t(\epsilon_t)$  according to variational approximation ( $l$  and  $t$  represent any relevant index for calculating all of the possible  $q_l(\mu_l)$  and  $q_t(\epsilon_t)$ ). The known parameters are represented with  $\Theta = \{\lambda_{1:R}^{-1}, \lambda^{-1}, \mu, \tau_{1:C}^{-1}, \tau^{-1}, \epsilon\}$ .

$$\begin{aligned}\ln q_l(\mu_l) &= E_{\substack{\epsilon_{1:C} \\ \mu_{1:R-l}}} [\ln p(S, \mu_{1:R}, \epsilon_{1:C} | \Theta)] + const \\ \ln q_t(\epsilon_t) &= E_{\substack{\mu_{1:R} \\ \epsilon_{1:C-t}}} [\ln p(S, \mu_{1:R}, \epsilon_{1:C} | \Theta)] + const\end{aligned}$$

To calculate the above we first calculate  $\ln q_l(\mu_l)$  which needs in return to first calculate  $p(S, \mu_{1:R}, \epsilon_{1:C})$  (Note that we choose indexes  $l$  and later  $t$  to denote a specific index so that it is not mixed with the indexes in the sum and product terms):

$$\begin{aligned}p(S, \mu_{1:R}, \epsilon_{1:C} | \Theta) &= p(S | \mu_{1:R}, \epsilon_{1:C}, \Theta) p(\mu_{1:R}, \epsilon_{1:C} | \Theta) = \\ &\prod_{c=1}^C \prod_{r=1}^R p(S_{rc} | \mu_r + \epsilon_c, \lambda_r^{-1} + \tau_c^{-1}) \prod_{r=1}^R p(\mu_r | \mu, \lambda^{-1}) \prod_{c=1}^C p(\epsilon_c | \epsilon, \tau^{-1})\end{aligned}$$

Notice that each  $\mu_r$  and each  $\epsilon_c$  are independent in the above expression, resulting in factors. Now we calculate  $\ln q_l(\mu_l)$  by taking the expectation over the log of the expression above:

$$\ln q_l(\mu_l) = -\frac{1}{2} E_{-\mu_l} \left[ \sum_{c=1}^C \sum_{r=1}^R \frac{(S_{rc} - (\mu_r + \epsilon_c))^2}{\lambda_r^{-1} + \tau_c^{-1}} + \frac{1}{\lambda^{-1}} \sum_{r=1}^R (\mu_r - \mu)^2 + \frac{1}{\tau^{-1}} \sum_{c=1}^C (\epsilon_c - \epsilon)^2 \right] + const$$

All the components of  $r$  in  $\sum_{r=1}^R$  are independent on  $\mu_l$  except  $r = l$ ; adding those independent terms in the above into constant results in:

$$\begin{aligned}& -\frac{1}{2} E_{\epsilon} \left[ \sum_{c=1}^C \frac{S_{lc}^2 - 2S_{lc}(\mu_l + \epsilon_c) + (\mu_l^2 + \epsilon_c^2 + 2\mu_l \epsilon_c)}{\lambda_l^{-1} + \tau_c^{-1}} + \lambda(\mu_l^2 + \mu^2 - 2\mu\mu_l) + \frac{1}{\tau^{-1}} \sum_{c=1}^C (\epsilon_c - \epsilon)^2 \right] \\ & + const\end{aligned}$$

The  $\mu_l$ -independent terms in the above expression can be pushed into constant resulting in (the expectation goes through the terms as below):

$$\begin{aligned}\ln q_l(\mu_l) &= -\frac{1}{2} E_{\epsilon} \left[ \sum_{c=1}^C \frac{-2S_{lc}(\mu_l) + (\mu_l^2 + 2\mu_l \epsilon_c)}{\lambda_l^{-1} + \tau_c^{-1}} + \lambda(\mu_l^2 - 2\mu\mu_l) \right] + const \\ &= -\frac{1}{2} E_{\epsilon} \left[ \left( \lambda + \sum_{c=1}^C \frac{1}{\lambda_l^{-1} + \tau_c^{-1}} \right) \mu_l^2 - 2\mu_l \left( \mu\lambda + \sum_{c=1}^C \frac{S_{lc} - \epsilon_c}{\lambda_l^{-1} + \tau_c^{-1}} \right) \right] + const \\ &= -\frac{1}{2} \left[ \mu_l^2 \left( \lambda + \sum_{c=1}^C \frac{1}{\lambda_l^{-1} + \tau_c^{-1}} \right) - 2\mu_l \left( \mu\lambda + \sum_{c=1}^C E_{\epsilon_c} \left[ \frac{S_{lc} - \epsilon_c}{\lambda_l^{-1} + \tau_c^{-1}} \right] \right) \right] + const \\ &= -\frac{1}{2} \left[ \mu_l^2 \left( \lambda + \sum_{c=1}^C \frac{1}{\lambda_l^{-1} + \tau_c^{-1}} \right) - 2\mu_l \left( \mu\lambda + \sum_{c=1}^C \frac{S_{lc} - E_{\epsilon_c}[\epsilon_c]}{\lambda_l^{-1} + \tau_c^{-1}} \right) \right] + const\end{aligned}$$

By completing the square, the above expression becomes in a form of Gaussian with mean

$m_l$  and precision  $p_l$ , where

$$p_l = \lambda + \sum_{c=1}^C \frac{1}{\lambda_l^{-1} + \tau_c^{-1}}$$

$$m_l = \frac{\mu\lambda + \sum_{c=1}^C \frac{S_{lc} - E_{\epsilon_c}[\epsilon_c]}{\lambda_l^{-1} + \tau_c^{-1}}}{p_l}$$

Similarly we can calculate  $\ln q_t(\epsilon_t)$  and so it results in a Gaussian with mean  $m_t$  and precision  $p_t$ , where

$$p_t = \tau + \sum_{r=1}^R \frac{1}{\tau_t^{-1} + \lambda_r^{-1}}$$

$$m_t = \frac{\epsilon\tau + \sum_{r=1}^R \frac{S_{rt} - E_{\mu_r}[\mu_r]}{\tau_t^{-1} + \lambda_r^{-1}}}{p_t}$$

Since we have approximated  $q(\mu_l)$ , we have R number of estimations each having similar result as in  $mean_l$  and  $precision_l$ . Having  $q(\epsilon_t)$  form, we have C number of estimations each having similar result as in  $mean_t$  and  $precision_t$ ; thus, we can calculate  $\prod_{r=1}^R q(\mu_r) \prod_{c=1}^C q(\epsilon_c)$ .

## References

- [1] David M. Blei. *Variational Inference*. <https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf>. [Online; accessed 20-November-2019]. 2011.
- [2] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [3] E Weisstein. *Second derivative test*. From *MathWorld—A Wolfram Web Resource*. 2004.